Am. J. Hum. Genet. 42:173–177, 1988

Alternative Bioassays of Kinship between Loci

N. E. Morton and D. Wu

Genetic Epidemiology Division, Memorial Sloan-Kettering Cancer Center, New York

Summary

In this study four asymptotically equivalent estimates of kinship are derived, in the general case and for kinship between multiallelic loci. Two estimates based on χ^2 agree closely, with the Shannon estimate giving the smaller variance. The PAH, GH, GM, and HBB systems conform to a recombinational model with an evolutionary size of ~4,000 and a ratio of recombination to physical distance of $\sim 1.4 \times 10^{-5}$ morgans/kb, as predicted on the basis of the genetic and physical lengths of the human genome. The INS and D11S12 systems have a much more rapid decline of kinship with physical distance, suggesting overlapping RFLPs (unrecognized allelism), recombinational hot spots, or selection. Sources of error in predicting kinship over small distances are discussed.

Introduction

Genetic relationship between individuals, populations, and loci is measured by kinship, which may be predicted on the basis of genealogy or migration and bioassayed on the basis of gene frequencies, discrete phenotypes, quantitative traits, clans, or surnames (Morton et al. 1971). Here we are concerned with alternative bioassays based on gene frequencies and with their validation vis-à-vis a known population structure that specifies isolation by distance, a cladogram of diverging populations, or (in this paper) a linear order of linked genes. The question to be answered is which formulation permits inference of the known structure with minimal error. Similarity metrics that do not estimate kinship have no simple genetic interpretation and will not be considered.

Theory

Let Q_k be the kth gene frequency in an array of populations, and let q_{ki} be the corresponding frequency in the ith population ($k = 1, \ldots, K$; $i = 1, \ldots, I$). Then the expected homozygosity in a cross between populations i and j with kinship φ_{ij} is $E(q_{ki}q_{kj}) = Q_k^2 + \varphi_{ij}Q_k(1 - Q_k)$. After summing and rearranging, the homozygosity estimate of kinship is

$$\varphi_{ij}(\mathrm{H}) = \frac{\sum_k q_{ki}q_{kj} - \Sigma Q_k^2}{1 - \Sigma Q_k^2}.$$

Weighting each element in the original equation by $1/Q_k$ before summing and rearranging yields the χ^2 estimate of kinship

$$\varphi_{ij}(\mathrm{C}) = \frac{\Sigma(q_{ki}q_{kj}/Q_k) - 1}{K - 1}.$$

For a codominant system this is proportional to Pearsonian χ^2 with observed frequency $\sqrt{q_{ki}q_{kj}}$ and expectation Q_k. The corresponding Shannon estimate of kinship based on the χ^2 likelihood ratio is

$$\varphi_{ij}(\mathrm{S}) = \frac{2\sum_k \sqrt{q_{ki}q_{kj}} \ln(\sqrt{q_{ki}q_{kj}}/Q_k)}{K - 1},$$

evaluated under the convention that $0 \ln 0 = 0$. These estimates are unbiased for $i \neq j$. If codominant gene frequencies in the ith population are based on a disomic locus in a sample of N_i individuals, an unbiased estimate of kinship in the population is

$$\varphi_{ii} = \frac{\varphi_{ij} - 1/2N_i}{1 - 1/2N_i},$$

Received February 13, 1987; final revision received June 8, 1987.

Address for correspondence and reprints: Dr. Newton E. Morton, Genetic Epidemiology Division, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021.

0002-9297/88/4201-0023$02.00

where φ_{ij} as an estimate of kinship in the sample is given by the above equations for $i = j$.

These formulas do not exhaust kinship estimates. A less direct approach begins with gene frequency deviations, $\Delta_{ki} = q_{ki} - Q_k$, and uses the Wahlund (W) principle to write the covariance $\Delta_{ki}\Delta_{kj} = Q_k(1 - Q_k)\varphi_{kij}$. Then mean kinship is

$$\varphi_{ij} = \sum_k w_k \varphi_{kij} / \Sigma w_k ,$$

where $w_k > 0$ is a weight. If we take $w_k = 1 - Q_k$, we obtain the χ^2 and Shannon estimates. However, taking $w_k = Q_k(1 - Q_k)$ gives the Wahlund estimate of kinship,

$$\varphi_{ij}(\mathrm{W}) = \frac{\sum_k (q_{ki} - Q_k)(q_{kj} - Q_k)}{1 - \Sigma Q_k^2} .$$

This has the same expectation as the homozygosity estimate but in practice is different, being always positive for $i = j$. Of course, small values can become negative after correction for sample size. If gene frequencies are replaced by surnames so that homozygosity becomes isonymy, the above equations remain valid except that φ must be divided by 4.

Similar formulas apply to kinship between loci (Morton and Simpson 1983). Then q_{rs} is the haplotype frequency for allele s at locus i and allele r at locus j, and $Q_{rs} = q_r q_s$ is the expected haplotype frequency under linkage equilibrium. If $r = 1, \ldots, R$ and $s = 1, \ldots, S$, the number of df (replacing $K - 1$) is $(R - 1)(S - 1)$. Formulas for φ_{ij} are given in table 1. If both loci are diallelic, φ_{ij} reduces to ρ^2, where ρ is both the standardized disequilibrium statistic and the correlation between loci (Hill and Robertson 1968). The χ^2 estimate has been used for pairs of diallelic loci by Chakravarti et al. (1984*b*). This is generalized to any number of alleles in table 1, where φ_{ij} is the mean value of ρ^2.

Sometimes a particular allele is selected at one locus. Let q_s be the conditional frequency of allele s at the other locus, and let $Q_s > 0$ be the corresponding marginal frequency in the population. This special case is also given in table 1.

Since kinship between loci is asymptotically a quadratic form, it has bias $1/n$ if based on a count of n haplotypes (Weir and Hill 1986). After being denoted with a prime symbol, the formulas in table 1 are transformed to

$$\varphi_{ij} = \frac{\varphi'_{ij} - 1/n}{1 - 1/n} .$$

If the physical distance between two loci is k kilobases and if the ratio of the recombination frequency

Table 1

Kinship between Loci

Estimate	φ_{ij}: General	φ_{ij}: Selected Allele
(*H*)	$\dfrac{\sum_r \sum_s q_{rs}^2 - \sum_r q_r^2 \sum_s q_s^2}{1 - \sum_r q_r^2 \sum_s q_s^2}$	$\dfrac{\sum_s q_s^2 - \sum_s Q_s^2}{1 - \sum_s Q_s^2}$
(C)	$\dfrac{\sum_r \sum_s (q_{rs}^2/q_r q_s) - 1}{(R - 1)(S - 1)}$	$\dfrac{\sum_s (q_s^2/Q_s) - 1}{S - 1}$
(*S*)	$\dfrac{2\sum_r \sum_s q_{rs} \ln (q_{rs}/q_r q_s)}{(R - 1)(S - 1)}$	$\dfrac{2\sum_s q_s \ln (q_s/Q_s)}{S - 1}$
(W)	$\dfrac{\sum_r \sum_s (q_{rs} - q_r q_s)^2}{1 - \sum_r q_r^2 \sum_s q_s^2}$	$\dfrac{\sum_s (q_s - Q_s)^2}{1 - \sum_s Q_s^2}$

(θ) to physical distance is $R = \theta/k$, the expected kinship between loci at equilibrium between drift and recombination is

$$\varphi = \frac{1}{4N_e\theta + 1} = \frac{1}{4N_eRk + 1},$$

where N_e is the evolutionary size of the population. Since N_e and R are unknown and R varies among systems, we set $4N_eR = C$ as the parameter to be estimated. On the basis of the noncentral χ^2 distribution, the information about φ is taken to be $W = (n - 1)(1 + Ck)/(Ck + 2n - 1)$. To estimate C we minimize the function $f = \Sigma W[\varphi - 1/(1 + Ck)]^2/2$ for m estimates of kinship and stable weights W, with φ constrained to the 0,1 interval. This gives the Newton-Raphson iteration in which weights stabilize as C converges: $C = C_0 + U/K \rightarrow \hat{C} > 0$, where $U = -\Sigma Wk[\varphi - 1/(1 + Ck)]/(1 + Ck)^2$; $K = \Sigma Wk^2/(1 + Ck)^4$; and the SE of $\hat{C}$ is $\sqrt{2f/(m - 1)K}$. These calculations are performed by the computer program HAPLOKIN (Appendix).

Results

Morton and Simpson (1983), Chakravarti et al. (1984*a*, 1984*b*), and Morton and Lew (1985) developed methods to map closely linked loci on the basis of haplotype frequencies and concluded that kinship between loci is the appropriate statistic. We have applied our formulas to these and other data sets (table 2).

Homozygosity kinship consistently exceeds the χ^2 and Shannon estimates. We do not understand this systematic difference between asymptotically equivalent methods. The variance of homozygosity kinship is notably high. As remarked by Morton et al. (1971), homozygosity is dominated by large gene frequencies, which are reciprocally weighted in efficient estimates. Therefore we shall not consider the homozygosity estimate further.

The Wahlund estimate is comparable to the homozygosity estimate and less reliable than the two estimates based on χ^2. This is to be expected, since the Wahlund weights discriminate against rare genes. Even for equal gene frequencies, extreme linkage disequilibrium leads to great discrepancy between the homozygosity and Wahlund estimates, on the one hand, and the χ^2 or Shannon estimates on the other. For example, if $q_{11} = q_{22} = .5$ and $q_{12} = q_{21} = 0$, then in large samples $\varphi(H) = \varphi(W) = 1/3$, $\varphi(C) = 1$, and $\varphi(S) = 2\ln 2$, which is censored to 1.

Shannon and χ^2 kinship agree extremely well, reflecting the convergence of Pearsonian and likelihood ratio χ^2 in large samples. The variances are acceptably small, with a tendency for the Shannon estimate to have a smaller variance. This agrees with the generally better performance of likelihood ratio χ^2 in small samples (Fisher 1922).

Discussion

A remarkable feature of the data is that systems highly polymorphic over sequences <30 kb show rapid decline of kinship with distance. There are problems in studying kinship over distances not much greater than the RFLPs themselves. Length polymorphism may induce substantial relative error in nominal distance, and the RFLPs may overlap (unrecognized allelism). If some haplotypes are missing in a large sample, is this due to linkage disequilibrium or allelism? In the latter event, the "loci" should be

Table 2

$\hat{C}$ ± SE for Four Kinship Bioassays

System (Source[a])	Maximum k	$\hat{C}$ ± SE: φ(H)		φ(W)		φ(C)		φ(S)		V_c/V_s
PAH (1)	96.0	.786	.200	1.379	.313	.240	.056	.219	.052	1.15
INS (2)	20.1	32.756	19.098	92.321	19.538	4.662	2.654	4.185	1.760	2.27
GH (3)	38.0	.830	.682	.813	.434	.114	.062	.088	.045	1.93
D11S12 (4)	6.9	31.319	78.420	138.564	359.702	6.137	9.951	6.878	9.888	1.01
GM (5)	136.0	.268	.081	.550	.228	.194	.077	.158	.068	1.27
HBB (6)	50.2	1.369	.798	1.245	.627	.240	.118	.182	.085	1.94

[a] 1, Chakraborty et al. 1987; 2, Chakravarti et al. 1986; 3, Chakravarti et al. 1984*b*; 4, Barker et al. 1984; 5, Migone et al. 1985; and 6, Pagnier et al. 1984.

pooled. Since allelic correlations are negative, the effect of such pooling is to raise the estimate of kinship and therefore lower the estimate of *C;* for example, the closely linked *Rsa*I and *Taq*I sites in INS showed no − − haplotype in 35 U.S. blacks and 86 Caucasians and Pima Indians, suggesting that they form an allelic series over a 2-kb region. Treating them as such reduces the estimate of *C* from 4.7 to 3.0 for χ^2 kinship and from 3.1 to 2.9 for Shannon kinship.

Chiasma frequencies suggest that the 3×10^6 kb of the haploid genome have a genetic length of 30 morgans in males. If the X chromosome and the greater θ in females are taken into account, this implies that *R* for the whole genome is $\sim 1.4 \times 10^{-5}$ morgans/kb. N_e in human populations of diverse origin is thought to be ~4,000, on the basis of weak evidence. These values are consistent with $C = 4N_eR = .224$. It is not plausible that N_e should vary among systems within a population. Therefore much larger values of *C,* as observed for INS and D11S12, suggest either failure of the model (e.g., maintenance of polymorphism by selection) or a locally elevated value of *R*. For the HBB system, a recombinational hot spot has been inferred in the 9.1-kb region 5′ to the normal β gene (Chakravarti et al. 1984*a*). This is not apparent in those haplotypes carrying the sickle cell hemoglobin gene that are considered here, perhaps reflecting their short evolutionary history.

Our goal in this study was not to describe or interpret variation in estimates of *C* but to compare alternative kinship bioassays. It is clear that the homozygosity approach is inferior to the two χ^2-based metrics, of which the Shannon estimate has smaller variance in this material. We therefore advocate this metric for determining kinship between either populations or loci whenever data are reduced to haplotype frequencies. Conventional kinship bioassay remains preferable when (1) data are reported as phenotype frequencies for factor-union systems, (2) there may be dominance, (3) the information matrix is of manageable size, and (4) the object is to estimate kinship between populations (Morton 1975).

Acknowledgments

This is Genetic Epidemiology Division paper 344. This work was supported by grant GM 36295 from the National Institutes of Health.

Appendix

HAPLOKIN

This procedure estimates kinship between loci on the basis of haplotype frequencies.

I. Assignment

Data, job, summary, and prolix files are required. A control-point (CP) file is optional.

II. Input

The job file has a major control—HK (A,1) (B,3) (C,2) (N,4,n)—specifying numeric fields for as many as 12 loci and for the haplotype count *N*, where *n* is the sample size. Blanks are read as zero. FM, ED, and CC controls are required. An ED control specifies allele codes for a particular locus, e.g., ED (A,1,2). Alleles not specified are pooled into a class designated 0. An optional MC control gives physical coordinates of loci, e.g., MC (0, 5.4, 23.1, 66.1, . . .), for which the distance between loci may be calculated as $k_{ij} = |MC_i - MC_j|$.

III. Analysis

Four estimates of kinship between loci are calculated (table 1). C_0 is assumed to be .25. If during iteration $C \leqslant 0$, it is set to $C/2$. Only pairs of loci with specified physical distance are used to estimate *C*. Convergence is declared if $U^2(m - 1)/2fK < 10^{-10}$.

IV. Output

The summary file gives file assignments; allele frequencies for each locus; values of *R, S,* and *K;* and the four kinship estimates for each of the $n(n - 1)/2$ pairs of *n* loci, where *R* and *S* are the numbers of nonzero allele frequencies at the two loci. In addition, $\hat{C}$, the variances, and the SEs are computed. The prolix file gives haplotype frequencies for each pair of loci and successive iterations for each estimate of *C*.

References

Barker, D., T. Holm, and R. White. 1984. A locus on chromosome 11p with multiple restriction site polymorphisms. Am. J. Hum. Genet. 36:1159–1171.

Chakraborty, R., A. S. Lidsky, S. P. Daiger, F. Guttler, S. Sullivan, A. G. Dilella, and S. L. C. Woo. 1987. Polymorphic DNA haplotypes at the human phenylalanine hydroxylase locus and their relationship with phenylketonuria. Hum. Genet. 76:40–46.

Chakravarti, A., K. H. Buetow, S. E. Antonarakis, P. G. Waber, C. D. Boehm, and H. H. Kazazian. 1984*a*. Nonuniform recombination within the human β-gene cluster. Am. J. Hum. Genet. **36**:1239–1258.

Chakravarti, A., S. C. Elbein, and M. A. Permutt. 1986. Evidence for increased recombination near the human insulin gene: implication for disease association studies. Proc. Natl. Acad. Sci. USA **83**:1045–1049.

Chakravarti, A., J. A. Phillips, K. H. Mellits, K. H. Buetow, and P. H. Seeburg. 1984*b*. Patterns of polymorphism and linkage disequilibrium suggest independent origins of the human growth hormone gene cluster. Proc. Natl. Acad. Sci. USA **81**:6085–6089.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. Lond. [A] **222**:309–368.

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**:226–231.

Migone, N., G. DeLange, A. Piazza, and L. L. Cavalli-Sforza. 1985. Genetic analysis of eight linked polymorphisms within the human immunoglobulin heavy-chain region. Am. J. Hum. Genet. **37**:1146–1163.

Morton, N. E. 1975. Kinship, information, and biological distance. Theor. Popul. Biol. **7**:246–255.

Morton, N. E., and R. Lew. 1985. Mapping genetic systems by the supratype method. Hum. Genet. **70**:231–235.

Morton, N. E., and S. P. Simpson. 1983. Kinship mapping of multilocus systems. Hum. Genet. **64**:103–104.

Morton, N. E., S. Yee, D. E. Harris, and R. Lew. 1971. Bioassay of kinship. Theor. Popul. Biol. **2**:507–524.

Pagnier, J., J. G. Mears, O. Dunda-Belkhodja, K. E. Schaefer-Rego, C. Beldjord, R. L. Nagel, and D. Labie. 1984. Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc. Natl. Acad. Sci. USA **81**:1771–1773.

Weir, B. S., and W. G. Hill. 1986. Nonuniform recombination within the human β-globin gene cluster (letter). Am. J. Hum. Genet. **38**:776–778.